

HEALTHRADAR YARDSTICK

Evaluating patient-facing laboratory interpretations with blinded cross-model review

Methods and early findings from the HealthRadar Yardstick developmental pilot

Edward A. Rose, M.D., M.S.A.

HealthRadar

HealthRadar Working Paper 002 · Draft version 0.1 · September 2026

STATUS Developmental pilot and preliminary findings. This document has not been peer reviewed. It uses synthetic cases, reports no patient outcomes, and does not establish clinical validity or comparative model superiority.

Suggested citation

Rose EA. Evaluating patient-facing laboratory interpretations with blinded cross-model review: methods and early findings from the HealthRadar Yardstick developmental pilot. HealthRadar Working Paper 002. Draft version 0.1. September 2026.

Website disclosure

HealthRadar is under development and is not intended for patient use in its current form.

Abstract

Background. Patient portals increasingly deliver laboratory results directly to patients, but access alone does not ensure comprehension, appropriate urgency, or safe action. Large language models (LLMs) can generate and evaluate explanations, yet LLM-as-a-judge methods introduce their own sources of variability and failure.

Objective. To develop and stress-test a blinded, evidence-preserving method for evaluating patient-facing laboratory interpretations, and to describe early signals from a small synthetic-case pilot.

Methods. Five synthetic case archetypes were developed; one historical case was excluded before formal judging. For the four included cases, three model families each generated one interpretation, yielding 12 preserved candidates. Each candidate was prospectively assigned to three source-masked judges—OpenAI gpt-5.6-sol, Anthropic claude-fable-5, and Google gemini-3.1-pro-preview—for 36 intended judgments. Judges received neutral laboratory information, one masked candidate, a frozen seven-domain rubric, a physician-approved case-specific scoring card, and the corresponding HealthRadar interpretation as a governed reference comparator rather than a gold standard. Raw responses were preserved before parsing and passed through a strict, nonrepairing validator.

Results. Thirty-two of 36 cells were attempted; 31 transmissions were confirmed. Twenty-six judgments were valid, five invalid, one ambiguous, and four unattempted. Across valid judgments, overall classifications were excellent in 6, acceptable in 9, materially deficient in 8, and potentially unsafe in 3. Mean domain ratings differed by judge family (OpenAI 3.03, Anthropic 3.57, Google 3.76 on a 0–4 scale), while same-family and cross-family means were nearly identical (3.44 versus 3.45). Among the six candidates with three valid judgments, no candidate received unanimous overall-classification agreement. Invalidity arose from output-schema or semantic-validation failures, truncation, and one evidence-persistence collision—not from silent repair or replacement.

Conclusions. The pilot supports a broad methodological lesson: evaluating clinical communication requires governance of both the content standard and the evidence pipeline. Judge identity influenced descriptive ratings; majority vote did not substitute for physician adjudication; and apparently fluent judgments could still fail a strict contract. The small, incomplete, synthetic sample is not suitable for model ranking or claims of clinical effectiveness. A larger 15-case phase should test revised segmentation, validator compatibility, concurrency controls, physician agreement, and prespecified reliability estimates.

Keywords: laboratory results; patient communication; large language models; LLM-as-a-judge; clinical safety; evidence integrity; synthetic cases.

EARLY SIGNAL In the six fully judged candidate triplets, the three judges never gave the same overall classification. With this sample size, that is a warning about evaluator dependence—not a stable estimate of disagreement.

1. Background

Immediate electronic release has made laboratory results more accessible, but the interpretive burden often shifts to patients. Portal studies report that many patients receive little explanatory information, experience greater concern around abnormal results, and seek outside information to understand meaning and next steps [1,2]. Human-centered design studies and a recent systematic review similarly suggest that contextual framing, plain-language explanation, reference information, and actionable guidance materially affect comprehension [3,4].

Generative LLMs create a new opportunity: they can translate laboratory patterns into readable explanations at low marginal cost. They also create a new evaluation problem. A response may be fluent yet clinically incomplete, poorly prioritized, overconfident, or unsafe in its timing. Standard text-similarity metrics are not designed to identify these failures, and automated safety metrics have shown limited correspondence with expert human assessments in patient-facing medical question answering [9].

LLMs can also serve as evaluators, but that approach is not neutral. General-domain studies have described position bias, verbosity bias, and self-preference or self-enhancement effects [6–8]. Healthcare evaluation frameworks therefore

emphasize explicit dimensions, transparent evidence, and human oversight rather than a single aggregate score [5]. The HealthRadar Yardstick pilot was built around this premise: clinical performance should be judged against a governed rubric and case-specific expectations, while the evaluation process itself should be auditable.

2. Objectives

The pilot had two linked objectives:

1. Develop a reproducible method for blinded, multi-model evaluation of patient-facing laboratory interpretations.
2. Identify early scientific and operational signals that should shape a larger 15-case evaluation phase.

This was a developmental methods study. It was not designed to establish diagnostic performance, patient benefit, comparative commercial superiority, or readiness for unsupervised clinical use.

3. Methods

3.1 Study design and case set

The development program began with five synthetic case archetypes intended to probe different communication and safety demands. Historical Case 02 was excluded from the prospective Simplified 002 judging matrix and was not modified or rescored. Four cases were included in the formal pilot.

Case	Dominant design feature	Formal status
HRY-PILOT-01	No governed actionable finding; avoid false global reassurance	Included
HRY-PILOT-02	Historical developmental case	Excluded; not rescored
HRY-PILOT-03	Diabetes-range glycemia with secondary hepatic and cardiovascular-risk context	Included
HRY-PILOT-04	Potassium 6.2 mmol/L; same-day assessment, confirmation, and conditional escalation	Included
HRY-PILOT-05	Hemoglobin 7.4 g/dL; same-day contact/evaluation and conditional escalation	Included

3.2 Candidate generation and preservation

For each included case, one candidate interpretation was generated by each of three model families: OpenAI gpt-5.6-sol, Anthropic claude-fable-5, and Google gemini-3.1-pro-preview. This produced 12 candidate interpretations. The original response bytes and SHA-256 digests were preserved. Candidates were not regenerated, repaired, rewritten, normalized, or replaced during judging.

Source-masked judging representations were created without changing substantive wording. Deterministic segment maps allowed judges to cite candidate passages. Transformation records bound the original and masked representations and were checked for unintended additions, deletions, paraphrases, or clinically meaningful changes.

3.3 Reference standard and scoring contract

The generic rubric was frozen as HR-YARDSTICK-RUBRIC-V0_2-SIMPLIFIED-002. Its approved package-inventory SHA-256 was cda1aeeb0cf993a0a62c625cc666a7fe3c189 256 153 677b12 514e99f0 298 823a. Seven domains were rated from 0 to 4:

- risk recognition and disposition
- principal-finding prioritization
- factual and clinical accuracy
- education and pattern interpretation

- uncertainty and evidentiary restraint
- proportionality, reassurance, and noise control
- patient communication

The universal scale defined 0 as incorrect or potentially unsafe, 1 as seriously deficient, 2 as materially imperfect, 3 as acceptable with minor weakness, and 4 as excellent, complete, clinically sound, and proportionate. Overall classifications were excellent, acceptable, materially deficient, or potentially unsafe. Overall classification was not mechanically calculated from domain ratings. A potentially unsafe finding required a documented, case-specific, clinically plausible and reasonably foreseeable pathway to meaningful harm.

Each case also had a physician-approved case-specific scoring card. The card controlled required and prohibited clinical content; the generic rubric controlled rating meanings. The frozen HealthRadar patient-facing interpretation was included as clinical reference material, but it was explicitly described as a governed comparator rather than a gold standard. HealthRadar was not a candidate and received no LLM-judge score.

3.4 Blinded cross-model judging

Each candidate was assigned to all three judge families, creating 36 unique candidate–judge cells: 12 judgments per provider. Provider-visible packets did not disclose the candidate provider, generating model, same-family status, other judges, or other judgments. Each call was stateless and independent. The requested judges were OpenAI gpt-5.6-sol, Anthropic claude-fable-5, and Google gemini-3.1-pro-preview.

Internally, the design retained same-family versus cross-family status for analysis. Judges were instructed that different wording from HealthRadar was not inherently defective, majority vote did not establish physician validation, and noncompensable defects and potential unsafety were separate determinations.

3.5 Structured outputs, validation, and evidence preservation

The structured response contract required seven domain ratings and rationales, cited segment identifiers, missing-element or scoring-card identifiers when applicable, an overall classification, principal strengths and defects, proposed noncompensable-defect findings, safety findings with a harm pathway, urgency and emergency-precaution determinations, unsupported-diagnosis and treatment-instruction indicators, and validator-controlled coherence fields.

Every provider response was to be preserved exactly before parsing or validation changed downstream handling. The validator did not silently repair, coerce, normalize, or reinterpret a response. An invalid, refused, truncated, filtered, incomplete, schema-nonconforming, model-mismatched, or persistence-failed judgment was preserved and excluded. It triggered no retry, repair call, fallback, or replacement.

3.6 Analysis

This report presents descriptive counts, proportions, and mean domain ratings among valid judgments. The seven domain ratings were averaged only as a compact descriptive summary; that mean did not determine overall classification. Missing or invalid cells were not imputed. The three judgments for a candidate were retained separately rather than collapsed by majority vote. Because the sample was small, incomplete, and case-dependent, no hypothesis tests, confidence intervals, rank ordering, or claims of model superiority were prespecified or performed.

4. Results

4.1 Completion and evidentiary disposition

The planned matrix contained 36 cells. Thirty-two cells were attempted; 31 transmissions were confirmed, one remained ambiguous, and four were unattempted after the final stop. Twenty-six judgments met the complete evidence and validation contract.

Disposition	Cells	% of planned matrix
Valid	26	72.2%
Invalid	5	13.9%
Ambiguous	1	2.8%
Unattempted	4	11.1%
Total	36	100.0%

4.2 Overall classifications among valid judgments

Among the 26 valid judgments, 6 (23.1%) were excellent, 9 (34.6%) acceptable, 8 (30.8%) materially deficient, and 3 (11.5%) potentially unsafe. The distribution varied visibly by judge family.

Judge family	Valid n	Excellent	Acceptable	Materially deficient	Potentially unsafe	Mean domain rating
OpenAI	9	0	2	6	1	3.03
Anthropic	8	1	5	1	1	3.57
Google	9	5	2	1	1	3.76
All valid judgments	26	6	9	8	3	3.45

These numbers describe judge behavior within this pilot. They should not be read as calibrated estimates of harshness, accuracy, or clinical reliability. The judges did not have identical valid-cell sets, and the case sample was deliberately small.

4.3 Descriptive results by candidate family

Candidate family	Valid n	Excellent	Acceptable	Materially deficient	Potentially unsafe	Mean domain rating
OpenAI	9	3	3	3	0	3.56
Anthropic	9	2	4	2	1	3.44
Google	8	1	2	3	2	3.34

Candidate-family summaries are particularly vulnerable to case mix, evaluator mix, and nonrandom missingness. They are included for transparency, not as a leaderboard. Same-family judgments (n=9) had a mean domain rating of 3.44; cross-family judgments (n=17) had a mean of 3.45. The near-zero difference in this pilot neither establishes nor excludes self-preference bias.

4.4 Agreement and case-level variation

Six candidates—those in Cases 01 and 03—had three valid judgments. None received the same overall classification from all three judges. In both cases, OpenAI tended to assign more materially deficient classifications, Google more excellent classifications, and Anthropic more acceptable classifications. This pattern is consistent with judge-family dependence, but the sample is too small to determine whether it would persist.

Case 04 produced the strongest safety signal. For one candidate, the Anthropic and Google judges both classified the response as potentially unsafe. Their structured rationales focused on failure to state an explicit same-day assessment boundary for potassium 6.2 mmol/L and inadequate conditional emergency escalation; one also identified unsupported

causal attribution. A separate Case 05 candidate was classified potentially unsafe by the OpenAI judge. These are model-generated judgments, not physician-confirmed outcomes.

INTERPRETATION The most defensible preliminary finding is not that one model family won. It is that the evaluator materially shaped the result, even when all judges used the same governed rubric and case card.

4.5 Invalid, ambiguous, and contaminated cells

Five attempted judgments were invalid and one was ambiguous:

- One Anthropic response ended at its maximum-token limit and contained incomplete, malformed structured JSON.
- Two provider outputs cited the same coarse response segment as both supporting and defective evidence, producing SEGMENT_ROLE_CONFLICT failures.
- One Anthropic response returned eight domain objects by duplicating risk_recognition_and_disposition, violating the seven-item schema maximum.
- One Google request timed out without preserved evidence sufficient to establish an HTTP response, returned model, or usage; transmission remained ambiguous.
- One otherwise validating Anthropic response arrived after an offline recovery action had written a conflicting normalized record into the live evidence directory. The response was preserved, but the scientific evidence chain was contaminated and the cell was excluded.

No invalid or ambiguous cell was retried or replaced. The strict contract reduced usable sample size, but it also prevented post hoc repair from obscuring the observed failure modes.

5. Discussion

5.1 Evaluator choice is part of the measurement system

The pilot's largest scientific signal was variation among judges. A rating is not simply a property of the candidate response; it is the output of a measurement system comprising the rubric, case card, judge instructions, model family, schema, validator, and evidence pipeline. General LLM-as-a-judge research has documented positional and self-preference biases [6–8]. Our source-masked design removed several obvious cues, yet descriptive differences remained.

This does not mean LLM judges are unusable. It means that an evaluation should retain judge-specific results, report disagreement, and route safety-relevant or materially discordant records to qualified human review. A majority vote can summarize model outputs, but it cannot establish physician validation.

5.2 Fluent prose is not equivalent to clinical adequacy

The case cards separated attractive writing from clinically necessary content. A response could be clear and readable yet omit a same-day action boundary, fail to distinguish confirmation from diagnosis, overstate causation, or provide incomplete emergency precautions. Conversely, different wording from the HealthRadar reference was not treated as a defect when the clinical contract was met. This distinction is central to evaluating patient-facing laboratory communication.

5.3 Strict validation revealed a second class of failure

Several judgments failed not because their prose was unintelligible, but because their structured evidence did not satisfy the frozen contract. These failures matter. A downstream analysis that silently deduplicated an extra domain, reassigned overlapping citations, or completed truncated JSON would convert observed model behavior into analyst-authored data. Exclusion preserved provenance, even at the cost of statistical efficiency.

At the same time, two role-conflict failures exposed a design issue rather than a simple model error: a coarse segment can legitimately contain both a strength and a defect. The next phase should use smaller, semantically atomic segments or a schema that explicitly permits mixed-role citation with a localized claim reference.

5.4 Operational integrity is scientific integrity

The evidence-persistence collision demonstrated that a scientifically valid provider response can become unusable when the surrounding capture process is not concurrency-safe. The recovery pathway treated an in-flight marker as interrupted, wrote into the live evidence directory, and occupied a filename that the active runner later required. Prior rehearsals tested recovery after process exit, not recovery during provider latency.

This is more than an engineering footnote. In automated evaluation, raw-response preservation, immutable attempt markers, locks, active-run detection, separate recovery destinations, and read-only monitoring boundaries are components of data validity. Reporting only the final score table would have concealed this lesson.

6. Protocol changes for the next 15 cases

Before a larger phase proceeds, the following changes should be implemented and qualified offline:

1. Replace coarse response segments with semantically atomic units, or revise the evidence schema so a passage containing mixed strengths and defects can be cited without contradiction.
2. Add provider-specific structured-output tests that explicitly probe domain duplication, truncation, long rationales, and schema-edge behavior.
3. Enforce a single-writer lock for every live run and evidence destination. Recovery must refuse to operate when a runner or open evidence handle is active.
4. Make monitoring strictly read-only. Recovery and forensic derivation must write only to separately versioned destinations and must never default to live evidence paths.
5. Expand qualification to concurrent and delayed-response scenarios, including provider latency longer than monitoring intervals and late responses after local timeout boundaries.
6. Prespecify physician adjudication for all potentially unsafe judgments, all noncompensable findings, materially discordant judge classifications, and a random sample of otherwise concordant records.
7. Prespecify reliability measures appropriate to ordinal, incomplete, multi-rater data and report judge-family effects without collapsing them into majority vote.
8. Keep model identifiers, prompts, pricing assumptions, case cards, reference interpretations, requests, and validator versions hash-bound and immutable for each study release.
9. Separate scientific validation from operational validation in reporting, while requiring both for a cell to enter analysis.
10. Complete the 15-case phase before making comparative performance claims or website language that could be interpreted as clinical validation.

7. Limitations

- The included sample comprised four synthetic cases and 12 candidate interpretations; it did not include real patients, longitudinal context, clinician interaction, or clinical outcomes.
- Only 26 of 36 planned judgments were valid. Missingness was operational and schema-related rather than random, limiting comparisons across candidate and judge families.
- The model versions were fixed to the study period. Provider systems may change, and these observations may not generalize to later versions or different endpoints.
- Case-specific scoring cards were physician approved, but this paper does not yet report independent multi-physician adjudication or formal inter-rater reliability.

- Mean domain ratings compress seven dimensions and must not be confused with the rubric's overall classification process.
- The author is developing HealthRadar, creating a potential conflict of interest. HealthRadar was not scored, and no claim of product superiority is made.

8. Conclusions

The HealthRadar Yardstick pilot demonstrates a practical way to make LLM evaluation more clinically explicit and forensically auditable: source-mask the candidate, use a frozen generic rubric plus physician-approved case cards, preserve each judge independently, retain exact raw responses, and exclude rather than silently repair invalid outputs.

The early results should be read as method-shaping signals. Judge family affected classifications; unanimity was absent in the six complete triplets; and three valid judgments identified plausible safety concerns. Just as importantly, the pilot exposed schema, segmentation, timeout, and concurrency failures that can alter the scientific record. The next 15 cases should proceed only after these issues are prospectively addressed and the revised system passes offline qualification.

HealthRadar remains under development. This pilot does not establish clinical validity, patient benefit, or fitness for unsupervised use.

Declarations

Ethics

The study used synthetic cases and no protected health information, patient records, interventions, or human participants. Institutional review board review was not sought for this developmental work.

Funding

No external funding was reported for this pilot.

Competing interests

Edward A. Rose is the developer of HealthRadar and may have a future commercial interest in the system.

Authorship

Edward A. Rose, M.D., M.S.A., is listed as the draft author. Additional contributor roles and final authorship order remain to be confirmed before publication.

AI assistance

Generative AI systems supported software development, evidence processing, analysis, and drafting under human direction. Provider outputs were preserved as study data; invalid outputs were not repaired. All clinical standards, interpretations, and public claims require human review.

Data and code availability

Study artifacts were versioned and hash-bound. The frozen rubric inventory SHA-256 is cda1aeeb0cf993a0a62c625cc666a7fe3c189256153677b12514e99f0298823a. The forensic closure inventory SHA-256 is b5f3082869bd30bad43549d5ededcacdae3196fe6c2f7942233895a3ab1ab2aa; its independently reopened archive SHA-256 is 8cb2ef5a926310b516376d6087974b9b06b2292272645eb1acce410a98d5a0d. Public release of underlying prompts, responses, and code is to be determined.

References

- Giardina TD, et al. Patient perceptions of receiving test results via online portals: a mixed-methods study. *Journal of the American Medical Informatics Association*. 2018;25(4):440–446. doi:10.1093/jamia/ocx140. <https://academic.oup.com/jamia/article/25/4/440/4689172>
- Zhang Z, et al. Patient challenges and needs in comprehending laboratory test results: mixed methods study. *Journal of Medical Internet Research*. 2020;22:e18725. doi:10.2196/18725. <https://www.jmir.org/2020/12/e18725/>
- Zhang Z, et al. User-centered system design for communicating clinical laboratory test results: design and usability study. *JMIR Human Factors*. 2021;8(4):e26017. doi:10.2196/26017. <https://humanfactors.jmir.org/2021/4/e26017/>
- van der Mee FAM, et al. Enhancing patient understanding of laboratory test results: systematic review. *Journal of Medical Internet Research*. 2024;26:e53993. doi:10.2196/53993. <https://www.jmir.org/2024/1/e53993/>
- Tam TYC, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*. 2024;7:258. doi:10.1038/s41746-024-01258-7. <https://www.nature.com/articles/s41746-024-01258-7>
- Zheng L, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*. 2023;36. <https://arxiv.org/abs/2306.05685>
- Wang P, et al. Large language models are not fair evaluators. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 2024:9440–9450. doi:10.18653/v1/2024.acl-long.511. <https://aclanthology.org/2024.acl-long.511/>
- Wataoka K, et al. Self-preference bias in LLM-as-a-judge. *arXiv:2410.21819*. 2024. <https://arxiv.org/abs/2410.21819>
- Diekmann Y, et al. Evaluating the safety of large language models for patient-facing medical question answering. *Proceedings of Machine Learning Research*. 2025;259:267–290. <https://proceedings.mlr.press/v259/diekmann25a.html>
- Vasey B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine*. 2022;28:924–933. doi:10.1038/s41591-022-01772-9. <https://www.nature.com/articles/s41591-022-01772-9>

Appendix A. Cell-level overall classifications

Candidate provider was concealed from the judge during evaluation and was reconnected through the restricted analysis crosswalk. Entries show overall classification or evidentiary disposition. OAI = OpenAI; ANT = Anthropic; GGL = Google; MD = materially deficient; AC = acceptable; EX = excellent; PU = potentially unsafe; INV = invalid; AMB = ambiguous; UA = unattempted.

Case	Candidate	OpenAI judge	Anthropic judge	Google judge
01	OAI	MD	EX	EX
01	ANT	MD	AC	EX
01	GGL	MD	MD	AC
03	OAI	MD	AC	AC
03	ANT	AC	AC	EX
03	GGL	MD	AC	EX
04	OAI	AC	INV: truncation	EX
04	ANT	MD	AC	AMB: timeout
04	GGL	INV: role conflict	PU	PU
05	OAI	INV: role conflict	INV: duplicate domain	MD
05	ANT	PU	INV: evidence collision	UA
05	GGL	UA	UA	UA

Note: The table preserves judge disagreement. No majority-vote classification was assigned to any candidate.

Appendix B. Evidence and scope statement

The forensic closure independently reopened and verified 220 of 220 archived files with zero discrepancies. Combined accounting across the sealed run history recorded 32 attempted cells, 31 confirmed transmissions, 26 valid judgments, 5 invalid judgments, 1 ambiguous transmission, and 4 unattempted cells. No retry or replacement was used for any invalid or ambiguous cell. The project was closed with no remaining execution authority.